frontiers in Neuroscience

ORIGINAL RESEARCH
published: 10 July 2019
doi: 10.3389/fnins.2019.00679

Intra-Scanner and Inter-Scanner Reproducibility of Automatic White Matter Hyperintensities Quantification

Chunjie Guo[1†], Kai Niu[2†], Yishan Luo[3], Lin Shi[3,4], Zhuo Wang[1], Meng Zhao[5], Defeng Wang[1,4], Wan'an Zhu[1], Huimao Zhang[1†] and Li Sun[5*†]*

[1] Department of Radiology, The First Hospital of Jilin University, Changchun, China, [2] Department of Otorhinolaryngology Head and Neck Surgery, The First Hospital of Jilin University, Changchun, China, [3] BrainNow Medical Technology Limited, Sha Tin, Hong Kong, [4] Department of Imaging and Interventional Radiology, The Chinese University of Hong Kong, Sha Tin, Hong Kong, [5] Department of Neurology and Neuroscience Center, The First Hospital of Jilin University, Changchun, China

OPEN ACCESS

Edited by:
Ching-Po Lin,
National Yang-Ming University, Taiwan

Reviewed by:
Yi Su,
Banner Alzheimer's Institute, United States
Xin Di,
New Jersey Institute of Technology, United States

***Correspondence:**
Huimao Zhang
huimaozhanglinda@163.com
Li Sun
sjnksunli@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: 11 January 2019
Accepted: 13 June 2019
Published: 10 July 2019

Citation:
Guo C, Niu K, Luo Y, Shi L, Wang Z, Zhao M, Wang D, Zhu W, Zhang H and Sun L (2019) Intra-Scanner and Inter-Scanner Reproducibility of Automatic White Matter Hyperintensities Quantification. Front. Neurosci. 13:679. doi: 10.3389/fnins.2019.00679

Objectives: To evaluate white matter hyperintensities (WMH) quantification reproducibility from multiple aspects of view and examine the effects of scan–rescan procedure, types of scanner, imaging protocols, scanner software upgrade, and automatic segmentation tools on WMH quantification results using magnetic resonance imaging (MRI).

Methods: Six post-stroke subjects (4 males; mean age = 62.8, range = 58–72 years) were scanned and rescanned with both 3D T1-weighted, 2D and 3D T2-weighted fluid-attenuated inversion recovery (T2-FLAIR) MRI across four different MRI scanners within 12 h. Two automated WMH segmentation and quantification tools were used to measure WMH volume based on each MR scan. Robustness was assessed using the coefficient of variation (CV), Dice similarity coefficient (DSC), and intra-class correlation (ICC).

Results: Experimental results show that the best reproducibility was achieved by using 3D T2-FLAIR MRI under intra-scanner setting with CV ranging from 2.69 to 2.97%, while the largest variability resulted from comparing WMH volumes measured based on 2D T2-FLAIR MRI with those of 3D T2-FLAIR MRI, with CV values in the range of 15.62%–29.33%. The WMH quantification variability based on 2D MRIs is larger than 3D MRIs due to their large slice thickness. The DSC of WMH segmentation labels between intra-scanner MRIs ranges from 0.63 to 0.77, while that for inter-scanner MRIs is in the range of 0.63–0.65. In addition to image acquisition, the choice of automatic WMH segmentation tool also has a large impact on WMH quantification.

Conclusion: WMH reproducibility is one of the primary issues to be considered in multicenter and longitudinal studies. The study provides solid guidance in assisting multicenter and longitudinal study design to achieve meaningful results with enough power.

KEY POINTS

- The intra-scanner and inter-scanner WMH reproducibility study in the same cohort.
- The best reproducibility was achieved by using 3D T2-FLAIR MRI under intra-scanner setting.
- There is a large variability in comparing WMH quantification results based on 2D T2-FLAIR MRI with those of 3D T2-FLAIR MRI.

Keywords: reproducibility of results, white matter, magnetic resonance imaging, brain, imaging, three-dimensional

INTRODUCTION

White matter hyperintensities, commonly found on T2-weighted T2-FLAIR brain MR images in the elderly, are associated with a number of neuropsychiatric disorders, including multiple sclerosis (MS) (Filippi et al., 2016),vascular dementia, Alzheimer's disease (AD) (Fazekas et al., 1996; Hirono et al., 2000), mild cognitive impairment (DeCarli et al., 2001), stroke (Fazekas et al., 1993), and Parkinson's disease (Marshall et al., 2006), and even in patients with primary mental disorders including mood disorders and schizophrenia spectrum disorders (Brown et al., 1995). Many studies have provided evidence that WMH have a strong impact on cognitive functioning (Gunning-Dixon and Raz, 2000) and they have been associated with impairment in a number of domains (Cees De Groot et al., 2000; Prins et al., 2004). WMH usually have a higher signal intensity compared to the normal-appearing white matter on FLAIR sequences and may appear iso- or hypointense on T1-weighted MR images. It can be measured quantitatively and non-invasively on large population samples and have been proposed as an intermediate marker, which could be used for the identification of new risk factors and potentially as a surrogate end point in clinical trials (Schmidt et al., 2004).

One challenging issue in studying WMH is the accurate and robust quantification and localization, given their variability and scattered spatial distribution. There are a number of automatic or semiautomatic methods and tools studying WMH segmentation and quantification, including thresholding method (Payne et al., 2002; Gibson et al., 2010; Simões et al., 2013), clustering methods (Admiraal-Behloul et al., 2005; Schmidt et al., 2012; Jain et al., 2015), and machine learning algorithms (Sweeney et al., 2014; López-Zorrilla et al., 2017; Rachmadi et al., 2018). While there are so many methods studying the accuracy of WMH segmentation and quantification, few studies examined the reproducibility of WMH quantification.

Accurate WMH quantification is of vital importance not only because it is associated with an increased risk of stroke, cognitive decline, dementia, and death, but also because their progression has been studied in association with cognitive decline, with increasing progression predicting a more rapid decline in global cognitive performance and executive function (Mungas et al., 2005; van den Heuvel et al., 2006; Kramer et al., 2007). In addition, WMH may also have a role as a surrogate marker to assess treatment efficacy. The impact of progression of WMH on stroke and dementia are also needed to help design therapeutic trials incorporating progression of WMH as an intermediate end point. In order to accurately observe the progression of WMH, the reproducibility of WMH measurement is of critical significance. The reliability of WMH quantification based on images acquiring from different scanners in multiple centers is of crucial importance in multi-center and follow-up studies. It is thought that a direct comparison of images or WMH quantities from different scanners in different centers may induce great variation, but no study examined the extent of this variation compared with within-center variability. The uncertain or lower reproducibility of WMH quantification across centers can contribute to a major concern for carrying out multicenter and longitudinal research, as well as clinical trials.

In this study, we carry out the study on the reproducibility of WMH quantification, which covers both intra-scanner and inter-scanner variability, 2D–3D magnetic resonance imaging (MRI) variability, MR system upgrade variability, and image processing tools variability in WMH quantification. The results of this study can provide great help and guidance in multicenter and longitudinal WMH study design.

Abbreviations: 95% CI, 95% confidence interval; CV, coefficient of variation; DSC, Dice similarity coefficient; FLAIR, fluid-attenuated inversion recovery; ICC, intra-class coefficient; SPM, statistical parametric mapping; WMH, white matter hyperintensities.

MATERIALS AND METHODS

Participants

Six post-stroke patients with last onset more than 6 months (4 male and 2 female; mean age = 62.8 years; range = 58–72 years) were prospectively recruited from the outpatient clinic at the Department of Neurology, the First Hospital of Jilin University, P.R. China. Exclusion criteria were cortical infarction > 1/3 hemisphere, severe neuropsychiatric disorders, and a history of traumatic brain injury or tumors. In addition, to exclude the confounding effect of edema, all the participants had been without treatment with dehydrating agent or steroid within 4 weeks before MRI scans. Based on visual assessment of WM lesions, Fazekas scale was assessed on all 3D T2-FLAIR images by an experienced radiologist (CJG), and the mean Fazekas scale score of each subject was recorded. The median Fazekas scale score of all participants was 2.2 (range 1–3) (Fazekas et al., 1987). The study was approved by the local ethics committee and written informed consent was obtained from all participants.

Image Acquisition

All participants were scanned within 12 h across four clinical MRI systems: MR1: 1.5-T Siemens Avanto (software: syngo MR B15); MR2: 1.5-T Siemens Avanto (software: syngo MR B17) (Siemens Healthcare, Erlangen, Germany); MR3: 3.0-T Philips Ingenia (Philips Healthcare, Best, the Netherlands), and MR4: 3.0-T Siemens Trio (Siemens Healthcare, Erlangen, Germany). 3D T1-weighted MRI sequence was obtained for assisting accurate WMH segmentation. 3D T2-FLAIR and 2D T2-FLAIR were acquired twice with repositioning in-between on each MRI system, resulting in a total of 16 T2-FLAIR volumes per participant. All the 2D T2-FLAIR parameters were from the default clinical sequences. The MRI acquisition parameters are detailed in **Table 1**.

Image Processing

Two fully automated WMH segmentation and quantification software were used for WMH segmentation and volumetric measurement. One is AccuBrain® (BrainNow Medical Technology Ltd.) and the other is lesion growth algorithm [16] as implemented in the Lesion Segmentation Toolbox (LST[1]). AccuBrain® is an automated brain segmentation and quantification software. It can segment a list of brain structures based on T1w MRI. Given additional T2-FLAIR MRI, it can also segment and quantify WMH (Shi et al., 2013). AccuBrain® segments T1w MRI and produces brain structure masks and tissue masks. Then, it coregisters T1w MRI with T2-FLAIR MRI and transforms the structure and tissue masks onto T2-FLAIR space. Using a set of morphological techniques, it extracts WMH on T2-FLAIR MRI and refines it using the transformed brain structure mask from T1w MRI. AccuBrain® is a cloud-based computing tool, which only requires MRI scans as input with no other tunable parameters.

LST is an open source toolbox of SPM used to segment T2 hyperintense lesions in FLAIR images. LST also relies on both T1w and T2-FLAIR MRI to segment WMH. It determines the three tissue classes of gray and white matter as well as cerebrospinal fluid from the T1w MRI and then uses the T2-FLAIR intensity distribution of each tissue class to detect outliers. The neighboring voxels are analyzed and assigned to lesions under certain conditions. This is done iteratively until no further voxels are assigned to lesions. Herein, the likelihood of belonging to WM or GM is weighed against the likelihood of belonging to lesions. We used the default parameters in LST toolbox, initial threshold: 0.3, MRF parameter: 1, and maximum iterations: 50.

[1]https://www.applied-statistics.de/lst.html

Reproducibility Analysis

To measure the reproducibility, several metrics, i.e., volume difference percentage, CV, DSC, and ICC, were computed. Volume difference percentage is defined as the percentage of quantified WMH volume difference between the two sequential scans of the average WMH volume value of the two scans:

$$\text{volume difference percentage} = \frac{|scan - rescan|}{(scan + rescan)/2} \times 100\%$$

TABLE 1 | MRI acquisition parameters.

	MR1	MR2	MR3	MR4
Manufacturer	Siemens	Siemens	Philips	Siemens
Model name	Avanto	Avanto	Ingenia	TrioTim
Station name	MEDPC26921	MRC25494	3FCD991	MRC35363
System version	syngo MR B15	syngo MR B17	R6.0.531.1	syngo MR B15
Field strength (T)	1.5	1.5	3	3
2D FLAIR				
Voxel size, mm^3	0.5 × 0.5 × 6.0	0.5 × 0.5 × 6.0	0.5 × 0.5 × 6.0	0.5 × 0.5 × 6.0
Number of slices	20	20	18	20
Repetition time (ms)	9,000	9,000	7,000	8,000
Echo time (ms)	99	103	93	93
Flip angle (°)	150	150	90	130
Voxel size, mm^3	1.0 × 1.0 × 1.0	1.0 × 1.0 × 1.0	1.0 × 1.0 × 1.0	1.0 × 1.0 × 1.0
Number of slices	176	176	344	176
Repetition time (ms)	7,500	7,500	4,800	7,500
Echo time (ms)	402	396	310	389
Flip angle (°)	120	120	90	120
3D T1WI				
Voxel size, mm^3	1.0 × 1.0 × 1.0	1.0 × 1.0 × 1.0	1.0 × 1.0 × 1.0	1.0 × 1.0 × 1.0
Number of slices	176	176	192	176
Repetition time (ms)	1,900	1,900	7.07	1,900
Echo time (ms)	3.37	3.37	3.19	2.96
Flip angle (°)	15	15	7	9

FLAIR, fluid-attenuated inversion recovery; T1WI, T1-weighted images.

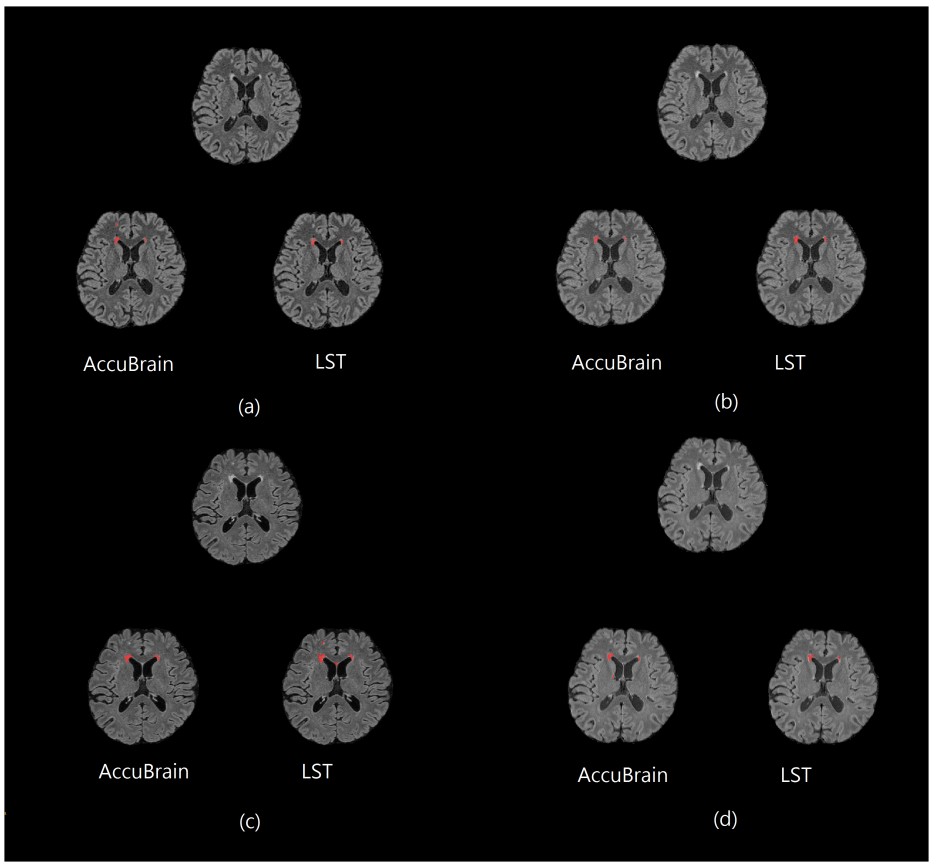

FIGURE 1 | One subject's 3D T2-FLAIR MR images from different scanners together with their WMH segmentation results (red overlay) using AccuBrain® and LST. **(a)** MR1; **(b)** MR2; **(c)** MR3; and **(d)** MR4.

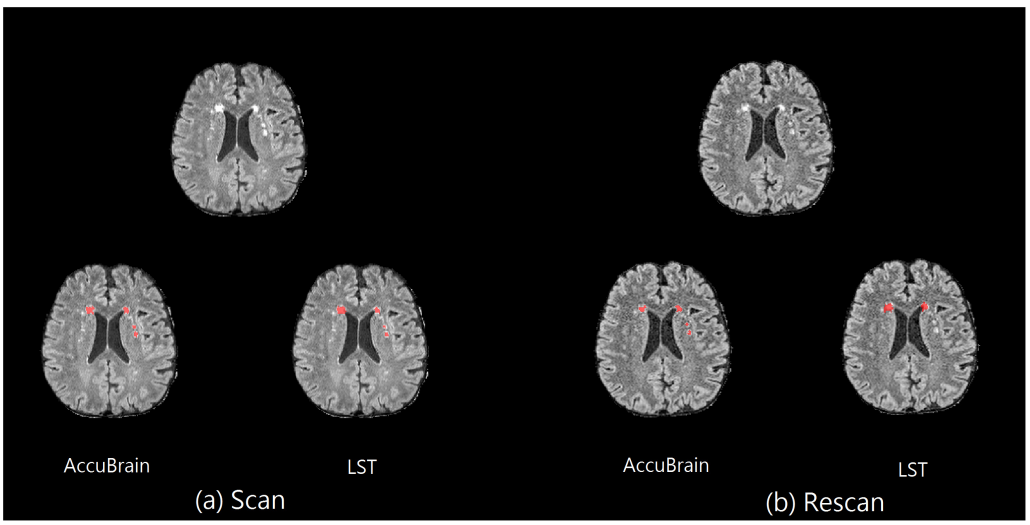

FIGURE 2 | Scan–rescan example on MR1. The corresponding T2-FLAIR MRI slice from a 3D T2-FLAIR MRI scan–rescan experiment on MR1 scanner, together with their WMH segmentation results using AccuBrain® and LST. **(a)** The first 3D T2-FLAIR scan. **(b)** Rescan with the subject's position change.

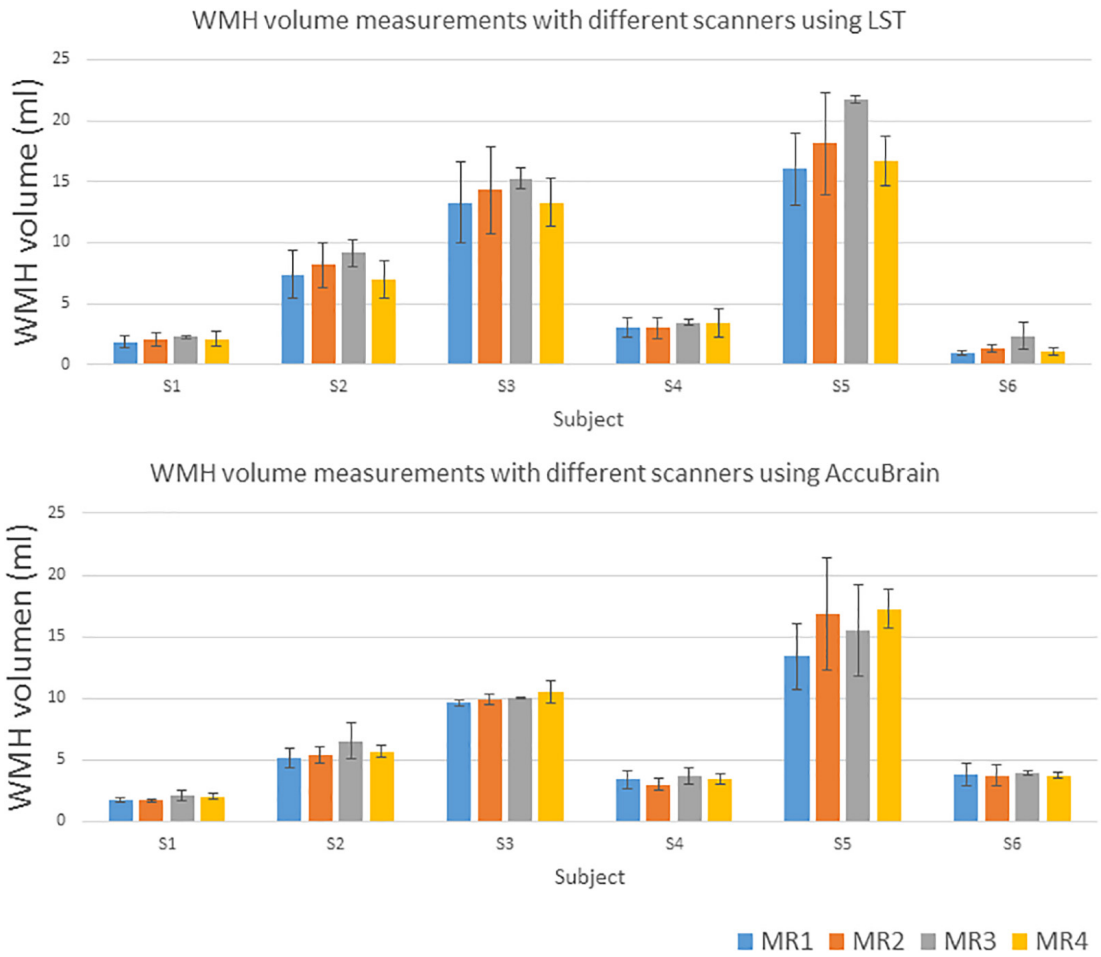

FIGURE 3 | WMH volume measurements using LST and AccuBrain® with MRIs from different scanners.

CV is defined as the ratio of the standard deviation to the mean of the multiple measurements and is expressed in percentages.

$$CV = \frac{\sigma}{\mu} \times 100\%$$

DSC is defined as the volume overlap of two segmentations:

$$DSC\,(A, B) = \frac{2\,(A \cap B)}{|A| + |B|}$$

In this study, we first aligned all the WMH results in the MR1 3D FLAIR MRI space, used the STAPLE algorithm (Warfield et al., 2004) to combine the WMH segmentation labels of all the scans, and created a fused label as reference label; each segmentation label was compared with the reference label in terms of DSC.

ICC was computed using two-way mixed method with 95% CIs in IBM SPSS Statistics 20 software.

The reproducibility of WMH quantification was assessed from four aspects.

Intra-Scanner Reproducibility

Each subject has a set of scanned 3D T1w, 2D T2-FLAIR, and 3D T2-FLAIR MRIs and re-scanned 2D T2-FLAIR and 3D T2-FLAIR MRIs on each MR scanner. The set of scan–rescan T2-FLAIR MRIs were used for examining within-scanner repeatability in a single center. The volume difference percentage, CV, DSC, and ICC between two sequential WMH measures using scan–rescan images were computed.

Inter-Scanner Reproducibility

The studying subjects were scanned with the same set of image sequences (3D T1w, 2D T2-FLAIR, and 3D T2-FLAIR) across four different scanners within 12-h interval. The inter-scanner variability was evaluated using CV, DSC, and ICC values of the same subject's different WMH measurements.

MR System Software Variability

Two of the four studying MRI scanners (MR1 and MR2) are the same MRI system from the same vendor (Siemens Avanto) but different in MRI system software (syngo MR B15 and syngo MR B17) and settled place. The effects of MR system upgrade

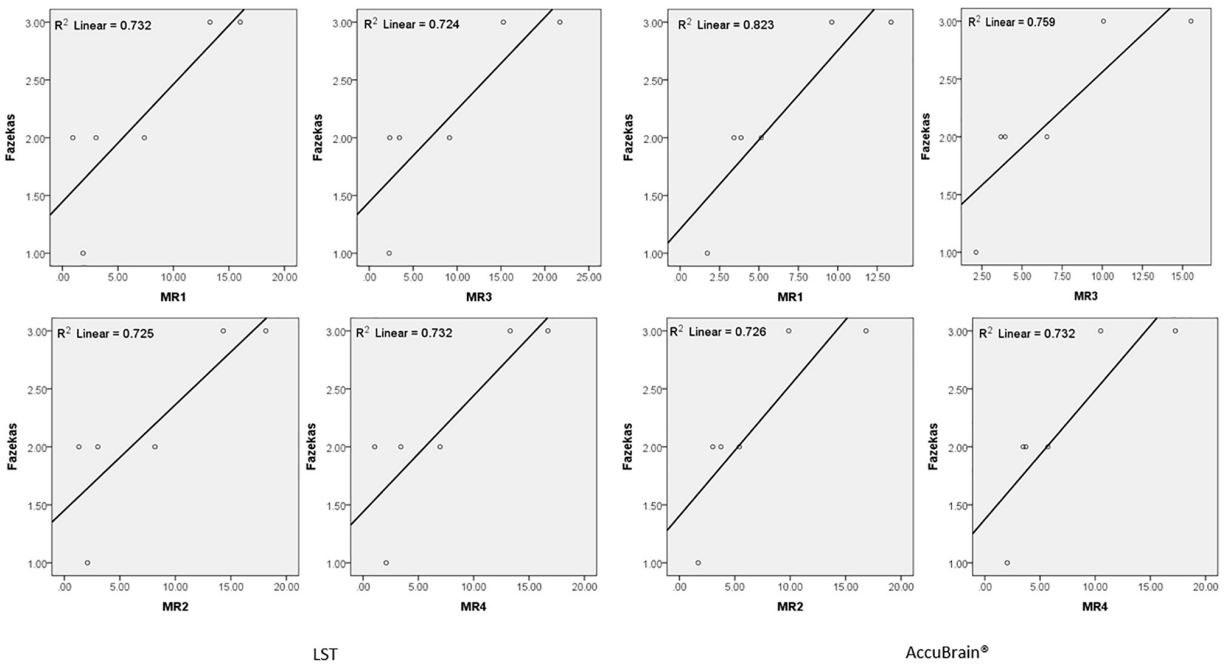

FIGURE 4 | The correlation between Fazekas score and WMH volume measurements using LST and AccuBrain® with MRIs from different scanners.

and examination place on WMH volume measurements were examined using this experiment.

2D and 3D T2-FLAIR Variability

As each subject was scanned using both 2D and 3D T2-FLAIR on the same scanner, the WMH volume measurement difference between 2D and 3D T2-FLAIR images was studied.

RESULTS

Segmentation

Figure 1 shows some representative axial slices of one subject's 3D T2-FLAIR MR images from different acquisitions and their corresponding automatic WMH segmentation results of the two software. It can be observed that the T2-FLAIR MRIs have a large variability in appearance across scanners, which brings great challenge in obtaining consistent WMH volumetric measurement.

In addition, on the same scanner, the subject's imaging position change can also have an impact on T2-FLAIR MRI appearance and WMH segmentation results. One example can be seen in **Figure 2**, where images from a 3D T2-FLAIR scan–rescan test are shown. Even if the same scanner and imaging parameters are used within a short time period, the T2-FLAIR MRIs look different in tissue and WMH contrast.

We quantified all the subjects' segmented WMH using both LST and AccuBrain® with all the MRIs from different scanners,

TABLE 2 | Intra-scanner WMH volume measurement reproducibility using different image processing software.

		Volume difference percentage (%) (mean ± std)		CV (%) (mean ± std)		DSC (mean ± std)		ICC (95% CI)	
3D	**Scanner**	**AccuBrain®**	**LST**	**AccuBrain®**	**LST**	**AccuBrain®**	**LST**	**AccuBrain®**	**LST**
	All	3.81 ± 2.97	4.20 ± 5.15	2.69 ± 2.10	2.97 ± 3.64	0.73 ± 0.06	0.74 ± 0.07	0.996 (0.992–0.998)	1 (0.999–1)
	MR1	4.35 ± 0.15	7.25 ± 0.11	3.07 ± 1.98	5.12 ± 6.34	0.70 ± 0.04	0.73 ± 0.10	0.995 (0.965–0.999)	0.999 (0.994–1)
	MR2	3.36 ± 0.10	2.17 ± 0.11	2.37 ± 2.80	1.53 ± 0.74	0.74 ± 0.09	0.70 ± 0.06	0.999 (0.996–1)	1 (0.998–1)
	MR3	4.13 ± 0.21	2.40 ± 0.14	2.92 ± 2.58	1.70 ± 1.38	0.77 ± 0.04	0.75 ± 0.05	0.992 (0.943–0.999)	1 (0.998–1)
	MR4	3.41 ± 0.11	4.97 ± 0.07	2.41 ± 1.19	3.51 ± 2.82	0.73 ± 0.05	0.77 ± 0.05	0.999 (0.989–1)	0.999 (0.996–1)
2D	All	7.35 ± 5.86	8.07 ± 8.06	5.19 ± 4.14	5.71 ± 5.70	0.68 ± 0.10	0.68 ± 0.12	0.969 (0.930–0.986)	0.997 (0.993–0.999)
	MR1	4.11 ± 2.98	6.48 ± 4.77	2.90 ± 2.11	4.58 ± 3.37	0.67 ± 0.12	0.63 ± 0.14	0.998 (0.989–1)	0.999 (0.990–1)
	MR2	9.68 ± 7.35	10.56 ± 13.32	6.84 ± 5.20	7.46 ± 9.42	0.68 ± 0.12	0.70 ± 0.15	0.980 (0.866–0.997)	0.998 (0.982–1)
	MR3	10.1 ± 6.39	10.76 ± 7.67	7.14 ± 4.52	7.60 ± 5.42	0.65 ± 0.08	0.70 ± 0.09	0.959 (0.741–0.994)	0.995 (0.968–0.999)
	MR4	5.51 ± 4.65	4.48 ± 2.31	3.89 ± 3.29	3.16 ± 1.63	0.69 ± 0.08	0.67 ± 0.09	0.981 (0.871–0.997)	0.998 (0.985–1)

FLAIR, fluid-attenuated inversion recovery; CV, coefficient of variation; DSC, Dice similarity coefficient; ICC, intra-class coefficient; 95% CI, 95% confidence interval.

TABLE 3 | Inter-scanner WMH volume measurement reproducibility using different image processing software.

	CV				DSC	ICC
	All scanners	95% CI of the difference	MR1 vs. MR2	95% CI of the difference	All scanners	All scanners
3D T2-FLAIR						
AccuBrain®	10.54 ± 4.09	(6.239–14.84)	5.01 ± 2.35	(2.538–4.485)	0.64 ± 0.082	0.985 (0.947–0.998)
LST	29.36 ± 24.37	(3.785–54.95)	6.97 ± 4.29	(2.466–11.47)	0.62 ± 0.129	0.950 (0.837–0.992)
2D T2-FLAIR						
AccuBrain®	11.49 ± 4.62	(6.646–16.34)	8.37 ± 5.41	(2.683–14.05)	0.63 ± 0.079	0.967 (0.888–0.995)
LST	10.89 ± 4.81	(5.836–15.95)	11.39 ± 7.61	(3.401–19.38)	0.65 ± 0.118	0.985 (0.949–0.998)

FLAIR, fluid-attenuated inversion recovery; CV, coefficient of variation; DSC, Dice similarity coefficient; ICC, Intra-class coefficient; 95% CI, 95% confidence interval.

TABLE 4 | Comparison of WMH quantification based on 2D and 3D T2-FLAIR.

	Intra-scanner CV		Inter-scanner CV	
	Mean ± std	95% CI of the difference	Mean ± std	95% CI of the difference
AccuBrain®	15.62 ± 8.73	(11.93–19.31)	17.17 ± 5.81	(11.07–23.27)
LST	24.19 ± 11.82	(19.19–29.18)	29.33 ± 15.01	(13.57–45.08)

CV, coefficient of variation; 95% CI, 95% confidence interval.

as shown in **Figure 3**. The WMH volumes range from 1 to 20 ml for different subjects. We quantified all the subjects' segmented WMH using both LST and AccuBrain® with all the MRIs from different scanners, as shown in **Figure 3**. The WMH volumes range from 1 to 20 ml for different subjects. S1–S6's mean Fazekas scale score is 1, 2, 3, 2, 3, and 2, respectively. The Pearson correlation between mean Fazekas scale score and LST quantified WMH volumes is 0.856 (MR1), 0.851 (MR2), 0.851 (MR3), and 0.856 (MR4), while the correlation is 0.907 (MR1), 0.852 (MR2), 0.871 (MR3), and 0.856 (MR4) with AccuBrain® quantified WMH volume, as shown in **Figure 4**.

Reproducibility

Table 2 shows the intra-scanner reproducibility results in the scan–rescan experiments. In general, the intra-scanner reproducibility results of different segmentation methods show relatively small differences with a mean volume difference percentage of 3.81% (3D) and 7.35% (2D) using AccuBrain®, and 4.20% (3D) and 8.07% (2D) using LST, and mean CV is 2.69% (3D) and 5.19% (2D) using AccuBrain®, and 2.97% (3D) and 5.71% (2D) using LST. The mean DSC is 0.73 (3D) and 0.68 (2D) using AccuBrain®, and 0.74 (3D) and 0.68 (2D) using LST. Comparatively, using 3D T2-FLAIR MRI brings more consistent WMH quantification results than using 2D-FLAIR MRI.

Table 3 compares the WMH volume measurement variability across different scanners. It shows that inter-scanner CV values are much larger than those of the intra-scanner experiment. Moreover, variability can also come from image processing software, where AccuBrain® has an average inter-scanner CV value of 10.54% (3D), while LST's inter-scanner CV value is 29.36% (3D) on average. In addition, MR1 and MR2 are two MRI scanners of the same type (Siemens Avanto) but with different versions of software installed and different rooms to be settled. There are still some variations between the quantifications on the two machines, but much smaller than that from different types of scanners.

In **Table 4**, the differences between 2D T2-FLAIR and 3D T2-FLAIR MRI were calculated and compared. It can be observed that in both intra-scanner and inter-scanner settings, the WMH volume measurement variations between 2D and 3D MRI are large and vary across different image processing tools.

DISCUSSION

In this study, we examined the reproducibility of WMH quantification. To achieve this, subjects with different levels of WMH load have undertaken MRI acquisitions (3D T1w, 2D and 3D T2-FLAIR sequences) across four different MRI scanners. On each scanner, a scan–rescan procedure was performed to examine intra-scanner variability, while the inter-scanner variability was tested across the four scanners. Meanwhile, the effect of software upgrade and settled place was examined using the same type of scanner but installed with different versions of software and in different examination rooms. In addition, comparison of WMH volume measurements between 2D and 3D T2-FLAIR was also made in both intra-scanner and inter-scanner setting.

In the intra-scanner experiments, it has shown that 3D T2-FLAIR MRIs generally achieved much better reproducibility than 2D T2-FLAIR MRIs regardless of image processing software, where the between-scan volume difference percentage is 3.81–4.20% for 3D T2-FLAIR and 7.35–8.07% for 2D T2-FLAIR. The larger variability of 2D scans indicates that the large slice thickness of 2D MRI scan can bring large variation in WMH volume measurement due to the irregular pattern of WMH across slices. An average 4% volume difference percentage was achieved in the scan–rescan procedure using 3D T2-FLAIR MRI. This implies that for a subject with low WMH load, e.g., 2 ml, a deviation of 0.08 ml may be induced on average with the same imaging setting; meanwhile, for a subject with

high WMH load, e.g., 10 ml, an average of 0.4 ml deviation may be induced. The scan–rescan reproducibility results can provide important clinical information in aiding doctor's further assessment or diagnosis.

There are several existing works studying the within-center reproducibility WMH volumetric measurement using a scan–rescan procedure in a single center. For example, de Boer et al. (2010) assessed whiter matter lesion segmentation reproducibility by comparing the automatic segmentations (by trained kNN method) of 30 subjects who were scanned twice within a short time interval; the mean CV is 5.87% using 3D T1w and 3D T2-FLAIR MRI. Another study assessed reproducibility of three automated segmentation pipelines for quantitative MRI measurement of brain white matter signal abnormalities (WMSA) on 30 subjects who were positioned and imaged twice within 30 min and achieved a range of 2.57–7.76% CV values using different pipelines (Wei et al., 2002). Ramirez et al. evaluated Lesion Explorer (LE), an MRI-derived tissue segmentation and brain region parcellation processing pipeline for obtaining intracranial tissue and subcortical hyperintensity volumetry in a short-term scan–rescan reliability test on 20 volunteers, with a reported intra-class correlation coefficient (ICC) of 0.9998 for subcortical hyperintensity measurement (Ramirez et al., 2013). In general, our reported intra-scanner CV values [CV: 2.69%, ICC: 0.996 (0.992–0.998) using AccuBrain® and CV: 2.97%, ICC: 1 (0.999–1) using LST] are close to the reported indices in a previous study. However, the previous studies mainly focused on 3D T2-FLAIR MRI. As the commonly used protocol in clinical practice, WMH quantification based on 2D T2-FLAIR MRI is also of great interest to clinicians. It is validated in our experiments that CV values of WMH quantification based on 2D MRI is in the range of 2.90–7.60% using different MRI scanners and processing software. In the inter-scanner experiments, the inter-scanner CV values (10.54% using AccuBrain® and 29.36% using LST) are around four to six times of the intra-scanner CV values (2.69% using AccuBrain® and 2.97% using LST). The large inter-scanner variability is mainly due to various T2-FLAIR MRI appearances resulting from different imaging parameters on different scanners. If the same scanner and imaging parameters are used, the difference can be smaller, where 5.01% and 6.97% CV value was achieved with AccuBrain® and LST, respectively, using the same Siemens Avanto scanner but with different versions of software and installation places. This variability is in the same level of intra-scanner variability, implying that machine software upgrade and installation place can have little impact on the measurement of WMH volume. However, it still suggests that centers should consider having some assessment or calibration for quality assurance and to calculate differences across time when scanner upgrade or replacement are considered.

In comparing quantification using both 2D and 3D T2-FLAIR, it has revealed that the variability is quite high under both intra-scanner and inter-scanner setting. 2D T2-FLAIR MRI is commonly accepted in clinical practice for diagnosis or assessment due to its relatively short acquisition time. However, WMH quantification results based on 2D MRI cannot be directly compared with the 3D MRI quantities, even with some resampling techniques, as it is easy to underestimate or overestimate WMH volume using 2D MRI.

Recommendations for multicenter WMH quantitative study:

(1) Acquire 3D T2-FLAIR MRIs using the same imaging parameters on the same scanner. Intra-scanner 3D T2-FLAIR reproducibility is much higher than others regardless of automatic quantification tools. Followed is the reproducibility of the same scanner but with upgraded software and resettled place. It indicates that in order to achieve the highest reproducibility, acquiring 3D T2-FLAIR MRIs in the same scanner is preferred in a multicenter study or a longitudinal study.

(2) WMH quantification on 2D T2-FLAIR MRIs is not comparable with that on 3D T2-FLAIR MRIs. Due to large slice thickness and irregular WMH pattern across slices, the variability of WMH volumetric measurement based on 2D T2-FLAIR MRI is much larger than 3D T2-FLAIR MRI. Although 2D T2-FLAIR MRI is commonly used in clinical practice, it is not preferable in a multicenter study or follow-up comparison. In particular, a direct comparison of quantitative results between 2D and 3D MRI can result in large deviation.

(3) WMH segmentation methods have a large impact on the quantification results and reproducibility. It can be observed that the scan–rescan reproducibility is relatively stable among different segmentation tools. However, the inter-scanner reproducibility is various among different tools. Choosing and comparing different image processing software is also an important issue in reliable WMH measurement.

(4) Before multicenter clinical trial is carried out, if different scanners are involved, protocol optimization and harmonization should be implemented first in each scanner. A reproducibility experiment with phantom or volunteer assessments for quality assurance is important to calculate differences in brain quantification.

(5) In a multicenter study, when images from different scanners have already been acquired, it is advised to (1) choose proper WMH quantification tools that are designed robust to scanner change and (2) use dedicated statistical models to adjust on scanner or use random-effects models.

Our study has several limitations. First, the subject number is not yet large enough for us to draw a statistically meaningful conclusion. Due to the long acquisition time for each subject to complete the whole procedure within 1 day, it is difficult to recruit many subjects in the current study. Second, MRI data acquisition was held in a single center. Although inter-scanner acquisitions were performed on different scanners and in different rooms to simulate multi-center study design, it is necessary in the future to launch multi-center reproducibility study based on a large population.

CONCLUSION

In conclusion, we compared WMH quantification reproducibility in different experimental settings. In general, the reproducibility

was the best when performing WMH segmentation on 3D MRI acquired by the same type of MRI scanner and same imaging parameters regardless of automatic segmentation tools. This study gives evidence on the extent of variability of WMH measurement across centers and can also aid in designing multicenter and longitudinal study to have enough power.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the First Hospital of Jilin University, China local ethics committee with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the First Hospital of Jilin University, China ethics committee.

AUTHOR CONTRIBUTIONS

CG, KN, YL, LSh, HZ, and LSu contributed to the study concept and design. CG, ZW, and MZ contributed to acquisition of the data. CG, KN, YL, LSh, DW, WZ, HZ, and LSu contributed to analysis and interpretation. CG, KN, YL, LSh, HZ, and LSu drafted the manuscript. All authors contributed to the critical revision of the manuscript for important intellectual content.

FUNDING

This study was supported by the grant provided by the Major Chronic Disease Program of the Ministry of Science and Technology of China (No. 2018YFC1312301), the Young Scholars Program of the National Natural Science Foundation of China (No. 81600923), and the General Program of Jilin Provincial Science and Technology Development of China (No. 2017C020).

REFERENCES

Admiraal-Behloul, F., van den Heuvel, D. M., Olofsen, H., van Osch, M. J., van der Grond, J., van Buchem, M. A., et al. (2005). Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *Neuroimage* 28, 607–617. doi: 10.1016/j.neuroimage.2005.06.061

Brown, F. W., Lewine, R. R. J., and Hudgins, P. A. (1995). White matter hyperintensity signals associated with vascular risk factors in schizophrenia. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 19, 39–45. doi: 10.1016/0278-5846(94)00102-N

Cees De Groot, J., de Leeuw, F. E., Oudkerk, M., van Gijn, J., Hofman, A., Jolles, J., et al. (2000). Cerebral white matter lesions and cognitive function: the Rotterdam scan study. *Ann. Neurol.* 47, 145–151.

de Boer, R., Vrooman, H. A., Ikram, M. A., Vernooij, M. W., Breteler, M. M., van der Lugt, A., et al. (2010). Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. *Neuroimage* 51, 1047–1056. doi: 10.1016/j.neuroimage.2010.03.012

DeCarli, C., Miller, B. L., Swan, G. E., Reed, T., Wolf, P. A., Carmelli, D., et al. (2001). Cerebrovascular and brain morphologic correlates of mild cognitive impairment in the national heart, lung, and blood institute twin study. *Arch. Neurol.* 58, 643–647. doi: 10.1001/archneur.58.4.643

Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I., and Zimmerman, R. A. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *Am. J. Roentgenol.* 149, 351–356. doi: 10.2214/ajr.149.2.351

Fazekas, F., Kapeller, P., Schmidt, R., Offenbacher, H., Payer, F., Fazekas, G., et al. (1996). The relation of cerebral magnetic resonance signal hyperintensities to Alzheimer's disease. *J. Neurol. Sci.* 142, 121–125. doi: 10.1016/0022-510X(96)00169-4

Fazekas, F., Kleinert, R., Offenbacher, H., Schmidt, R., Kleinert, G., Payer, F., et al. (1993). Pathologic correlates of incidental MRI white matter signal hyperintensities. *Neurology* 43, 1683–1683. doi: 10.1212/wnl.43.9.1683

Filippi, M., Rocca, M. A., Ciccarelli, O., De Stefano, N., Evangelou, N., Kappos, L., et al. (2016). MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *Lancet Neurol.* 15, 292–303. doi: 10.1016/S1474-4422(15)00393-2

Gibson, E., Gao, F., Black, S. E., and Lobaugh, N. J. (2010). Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T. *J. Magn. Reson. Imaging* 31, 1311–1322. doi: 10.1002/jmri.22004

Gunning-Dixon, F. M., and Raz, N. (2000). The cognitive correlates of white matter abnormalities in normal aging. *Quant. Rev.* 14:224. doi: 10.1037/0894-4105.14.2.224

Hirono, N., Kitagaki, H., Kazui, H., Hashimoto, M., and Mori, E. (2000). Impact of white matter changes on clinical manifestation of Alzheimer's disease: a quantitative study. *Stroke* 31, 2182–2188. doi: 10.1161/01.str.31.9.2182

Jain, S., Sima, D. M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., et al. (2015). Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *Neuroimage* 8, 367–375. doi: 10.1016/j.nicl.2015.05.003

Kramer, J. H., Mungas, D., Reed, B. R., Wetzel, M. E., Burnett, M. M., Miller, B. L., et al. (2007). Longitudinal MRI and cognitive change in healthy elderly. *Neuropsychology* 21, 412–418. doi: 10.1037/0894-4105.21.4.412

López-Zorrilla, A., de Velasco-Vázquez, M., Serradilla-Casado, O., Roa-Barco, L., Graña, M., Chyzhyk, D., et al. (2017). "Brain White Matter Lesion Segmentation with 2D/3D CNN," in *Natural and Artificial Computation for Biomedicine and Neuroscience*, eds J. Ferrández Vicente, J. Álvarez-Sánchez , F. de la Paz López, cpsfnmMoreo Jcpefnm Toledo, and H. Adeli (Berlin: Springer International Publishing), 394–403.

Marshall, G. A., Shchelchkov, E., Kaufer, D. I., Ivanco, L. S., and Bohnen, N. I. (2006). White matter hyperintensities and cortical acetylcholinesterase activity in parkinsonian dementia. *Acta Neurol. Scand.* 113, 87–91. doi: 10.1111/j.1600-0404.2005.00553.x

Mungas, D., Harvey, D., Reed, B. R., Jagust, W. J., DeCarli, C., Beckett, L., et al. (2005). Longitudinal volumetric MRI change and rate of cognitive decline. *Neurology* 65, 565–571. doi: 10.1212/01.wnl.0000172913.88973.0d

Payne, M. E., Fetzer, D. L., MacFall, J. R., Provenzale, J. M., Byrum, C. E., Krishnan, K. R., et al. (2002). Development of a semi-automated method for quantification of MRI gray and white matter lesions in geriatric subjects. *Psychiat. Res. Neuroim.* 115, 63–77. doi: 10.1016/S0925-4927(02)00009-4

Prins, N. D., van Straaten, E. C., van Dijk, E. J., Simoni, M., van Schijndel, R. A., Vrooman, H. A., et al. (2004). Measuring progression of cerebral white matter lesions on MRI. *Vis. Rat. volumetr.* 62, 1533–1539. doi: 10.1212/01.wnl.0000123264.40498.b6

Rachmadi, M. F., Valdés-Hernández, M. D. C., Agan, M. L. F., Di Perri, C., and Komura, T. (2018). Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. *Comput. Med. Imaging Graph.* 66, 28–43. doi: 10.1016/j.compmedimag.2018.02.002

Ramirez, J., Scott, C. J. M., and Black, S. E. (2013). A short-term scan–rescan reliability test measuring brain tissue and subcortical hyperintensity volumetrics obtained using the lesion explorer structural MRI processing pipeline. *Brain Topogr.* 26, 35–38. doi: 10.1007/s10548-012-0228-z

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., et al. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in

multiple sclerosis. *Neuroimage* 59, 3774–3783. doi: 10.1016/j.neuroimage.2011.11.032

Schmidt, R., Scheltens, P., Erkinjuntti, T., Pantoni, L., Markus, H. S., and Wallin, A. (2004). White matter lesion progression. *Surrog. Endpoint Trials Cereb. Small Vessel Dis.* 63, 139–144. doi: 10.1212/01.wnl.0000132635.75819.e5

Shi, L., Wang, D., Liu, S., Pu, Y., Wang, Y., Chu, W. C., et al. (2013). Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction. *J. Neurosci. Methods* 213, 138–146. doi: 10.1016/j.jneumeth.2012.12.014

Simões, R., Mönninghoff, C., Dlugaj, M., Weimar, C., Wanke, I., van Cappellen van Walsum, A. M., et al. (2013). Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images. *Magn. Reson. Imaging* 31, 1182–1189. doi: 10.1016/j.mri.2012.12.004

Sweeney, E. M., Vogelstein, J. T., Cuzzocreo, J. L., Calabresi, P. A., Reich, D. S., Crainiceanu, C. M., et al. (2014). A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI. *PLoS One* 9:e95753. doi: 10.1371/journal.pone.0095753

van den Heuvel, D. M. J., ten Dam, V. H., de Craen, A. J., Admiraal-Behloul, F., Olofsen, H., Bollen, E. L., et al. (2006). Increase in periventricular white matter hyperintensities parallels decline in mental processing speed in a non-demented elderly population. *J. Neurol. Neurosurg. Psychiatry* 77, 149–153. doi: 10.1136/jnnp.2005.070193

Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* 23, 903–921. doi: 10.1109/TMI.2004.828354

Wei, X., Warfield, S. K., Zou, K. H., Wu, Y., Li, X., Guimond, A., et al. (2002). Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. *J. Magn. Reson. Imaging* 15, 203–209. doi: 10.1002/jmri.10053

Conflict of Interest Statement: LSh was the director of BrainNow Medical Technology Limited. YL was an employee of BrainNow Medical Technology Limited, which developed AccuBrain® used in this manuscript.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.